Multi-Dataset Collection Research Scenario – GPCP and TMPA

George J. Huffman

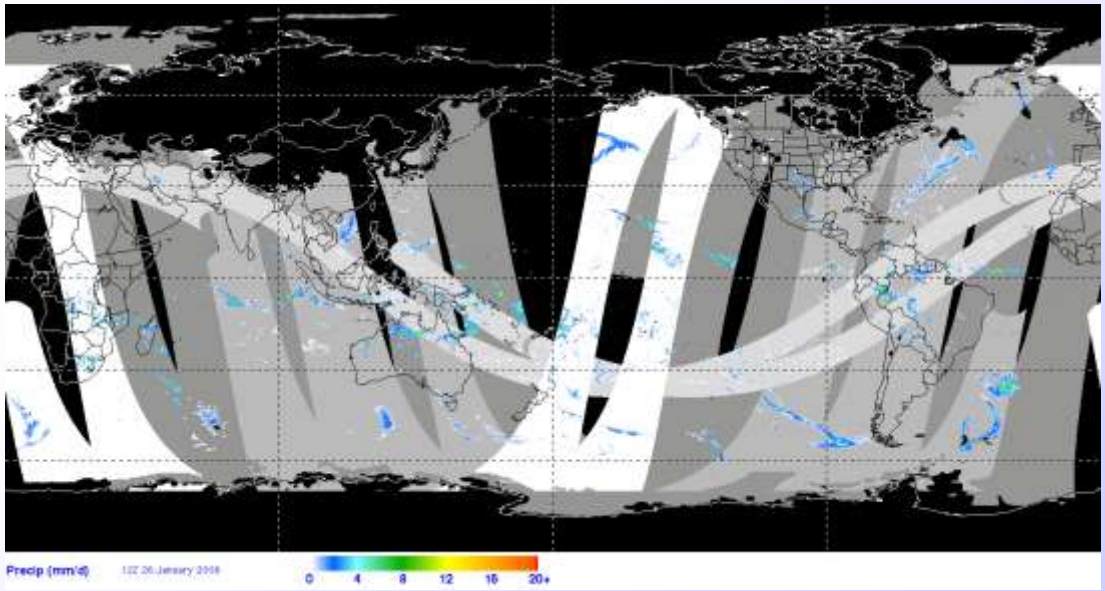
NASA/GSFC Laboratory for Atmospheres
Science Systems and Applications, Inc.

Colleagues: David T. Bolvin, Eric J. Nelkin

White, light-, medium-, dark-gray are AMSR, TMI, SSMI, AMSU swaths in TMPA

- note drop-outs in N. Hemisphere wintertime land

What does it take to add the next satellite to this picture?



General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build reading code

(Re)grid to our analysis grid

Sort out dataset quirks

Use the data

Update the local archive for new data, new versions, history of data faults

Compare/contrast with other versions/sources of same data

Other Considerations for CEWIS

Documentation

Which data?

Current case study - SSMIS is being integrated into both

Global Precipitation Climatology Project

- 1979-present, 90°N-S; 2.5° monthly and pentad
- 1997-present, 90°N-S; 1° daily

TRMM Multi-Satellite Precipitation Analysis

- 1998-present, 50°N-S; 0.25° monthly, 3-hr, and 3-hr real-time

Microwave Data in the GPCP and TRMM Combinations

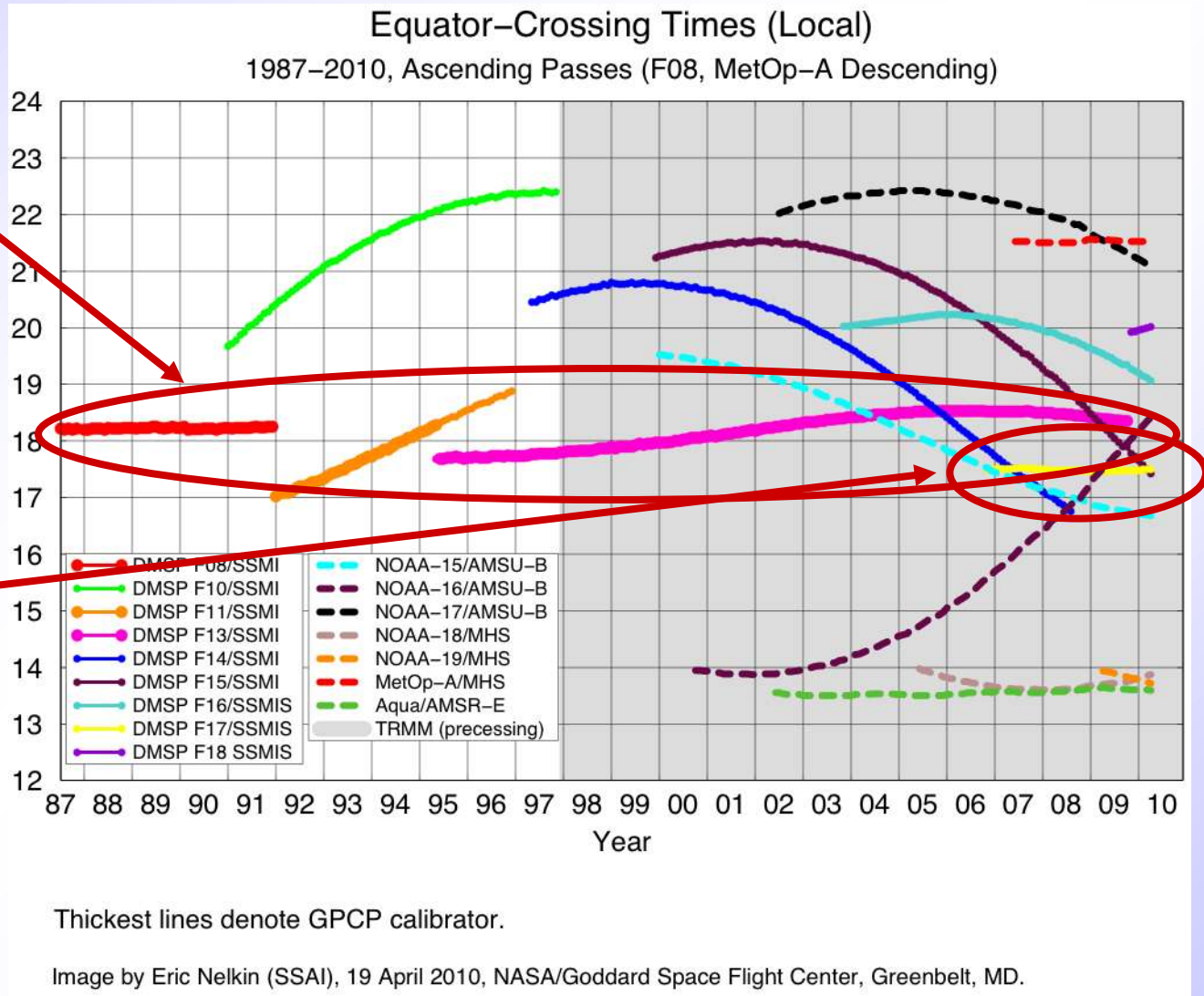
GPCP and TMPA both use microwave data

- GPCP uses only the 6 a.m./p.m. SSMI
- TMPA uses "everything"

All SSMI's are now done

DMSP F16, F17, F18 carry SSMIS

- F17 closest to 6 a.m./p.m. for GPCP
- TMPA needs them all



General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build reading code

(Re)grid to our analysis grid

Sort out dataset quirks

Use the data

Update the local archive

Compare/contrast with other versions/sources of same data

SSMIS the designated successor to SSMI

General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build reading co

(Re)grid to our analysis grid

Sort out dataset quirks

Use the data

Update the local archive

Compare/contrast with other versions/sources of same data

But, there have been lots of issues with calibration

- whose re-calibrated dataset should be used?
- is the algorithm of choice (GPROF2008) ready?

General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build reading code

(Re)grid to our analysis grid

Sort out dataset quirks

Use the data

Update the local archive

Compare/contrast with other

Major format choices

- big- or little-endian
- scaled integer vs. floats
- ASCII, flat binary, formatted (local, NetCDF, HDF)
- variable names/definitions
  - units
  - previous standardization effort fell flat!
- date/time representation
- COTS application treatment of filename suffixes (.doc, .txt, …)

When GPROF2008 is ready, the format is not like any of the previous GPROF's - it's "simple" binary.

General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build reading code

(Re)grid to our analysis grid

Sort out dataset quirks

Use the data

Update the local archive

Compare/contrast with other

We tend to modularize read routines for reuse and maintainability.

We tend to code readers (in Fortran) as opposed to using COTS applications – we're control freaks.

General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build reading code

(Re)grid to our analysis grid

Sort out dat

Use the data

Update the local archive

Compare/contrast with other

We try to pull Level 2 (swath) data when possible and grid it ourselves; again, it's control

- small (compared to gridbox) footprints are "forward gridded" - footprints assigned to whatever gridbox contains their center
- larger footprints are "backward gridded" - parceled out proportional to areas in each gridbox
- there are several major details in grid style
  - grid centers or edges?
  - CED, equal-area, or other?
  - full or partial globe?
  - row- or column-major?
  - start at north or south, Dateline or Prime Meridian?

General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build readi

(Re)grid to our analys

Sort out dataset quirks

Use the data

Update the local archive

Compare/contrast with other

This is the messiest part – with what do we have to cope to use the data?

- special values for “missing” and other special situations
- missing-filled vs. size-zero vs. absent files when granule is entirely without data
- available metadata, and its representation (in file name, header, ancillary file; positional, keyword)
- start/end padding scans
- partially/totally redundant granules
- variations in skill by region and/or period
- typical errors – always an adventure!
  - datetime errors, time/orbit mistmatches, navigation errors
  - sensor: hot/cold load drift, solar heating, sun glint, scan position biases

General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build reading code

(Re)grid to our analysis grid

Sort out dataset quirks

Use the data

Update the local archive

Compare/contrast with other versions/sources of same data

Actually do the computations and evaluate the results

General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build reading code

(Re)grid to our analysis grid

Sort out dataset quirks

Use the data

Update the local archive

Compare/contrast with other versions/sources of same data

New data, or new versions of the data

History of data faults, presumably accounted for by our processing

General Approach

Hear about the data

Get samples of the data

Figure out the format

Modify/adapt/build reading code

(Re)grid to our analysis grid

Sort out dataset quirks

Use the data

Update the local archive

Compare/contrast with other versions/sources of same data

Helps us understand dataset performance

Analysis with alternative datasets and comparison with other analyses builds confidence in the result

Other Considerations for CEWIS

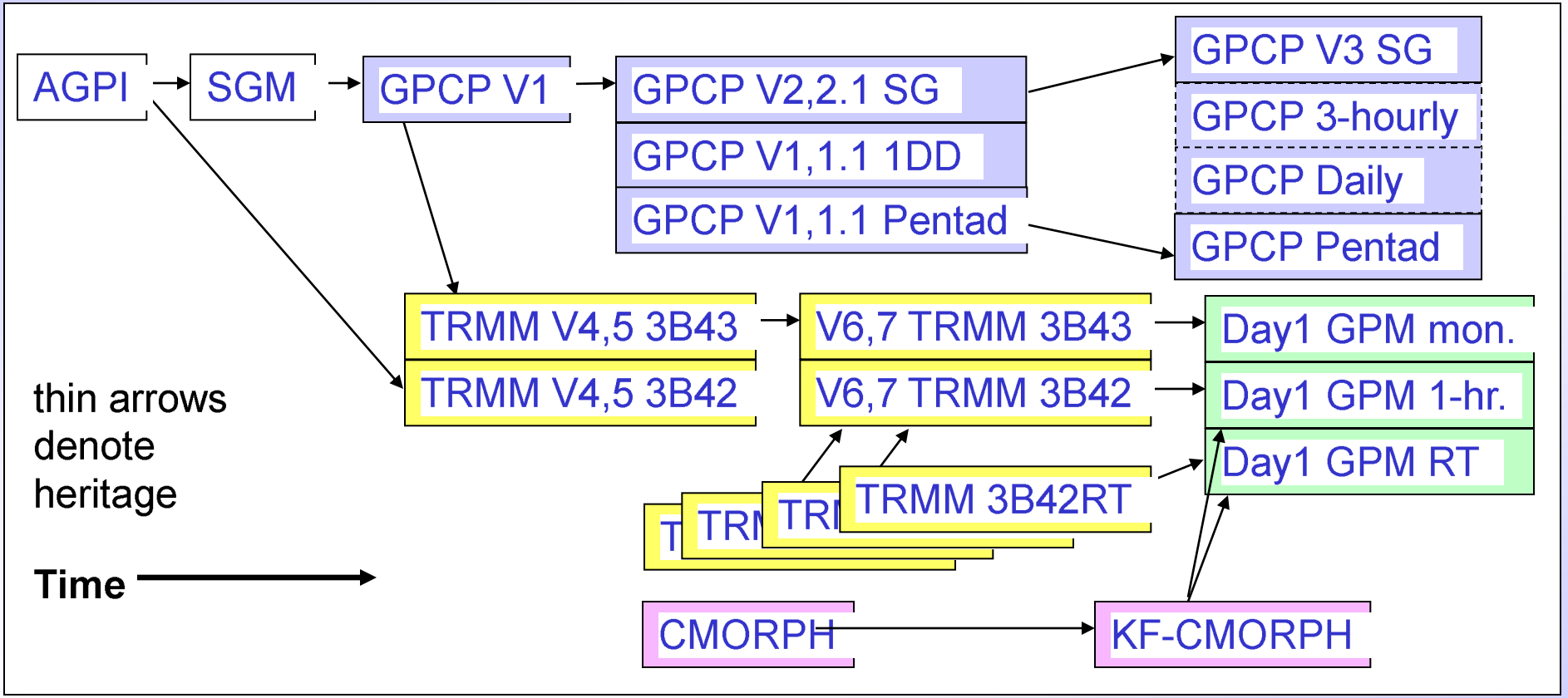
Documentation

- (pointers to) original paper(s), tech document, README
- clear “what’s different” README
- sample reading software, scripts, and/or macros
- “known errors and issues” log
- satellite and algorithm history
  - start/stop date/times
  - version names
  - extent of reprocessing
  - changes of time/space coverage and resolution

Other Considerations for CEWIS (cont.)

Documentation (cont.)

- example of GPCP and TMPA histories



Other Considerations for CEWIS (cont.)

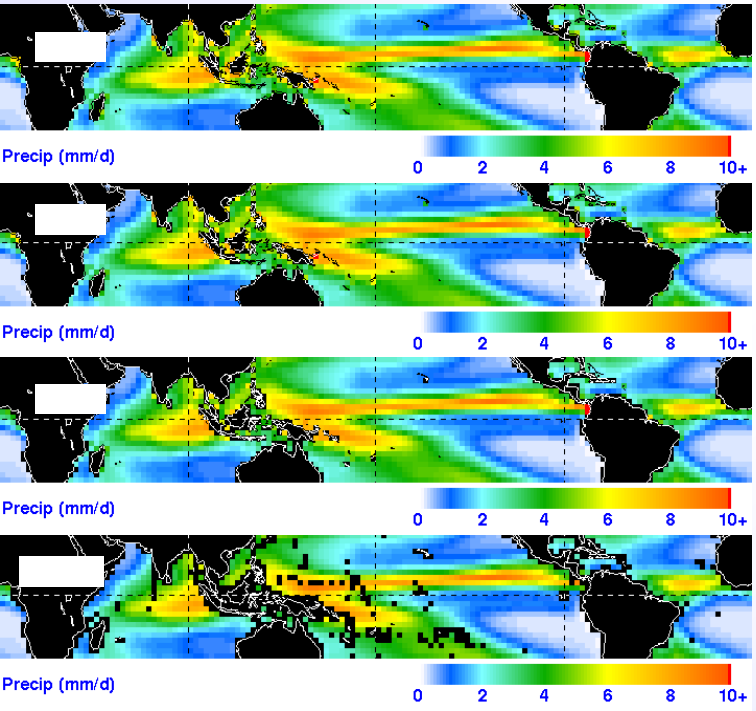
Which data?

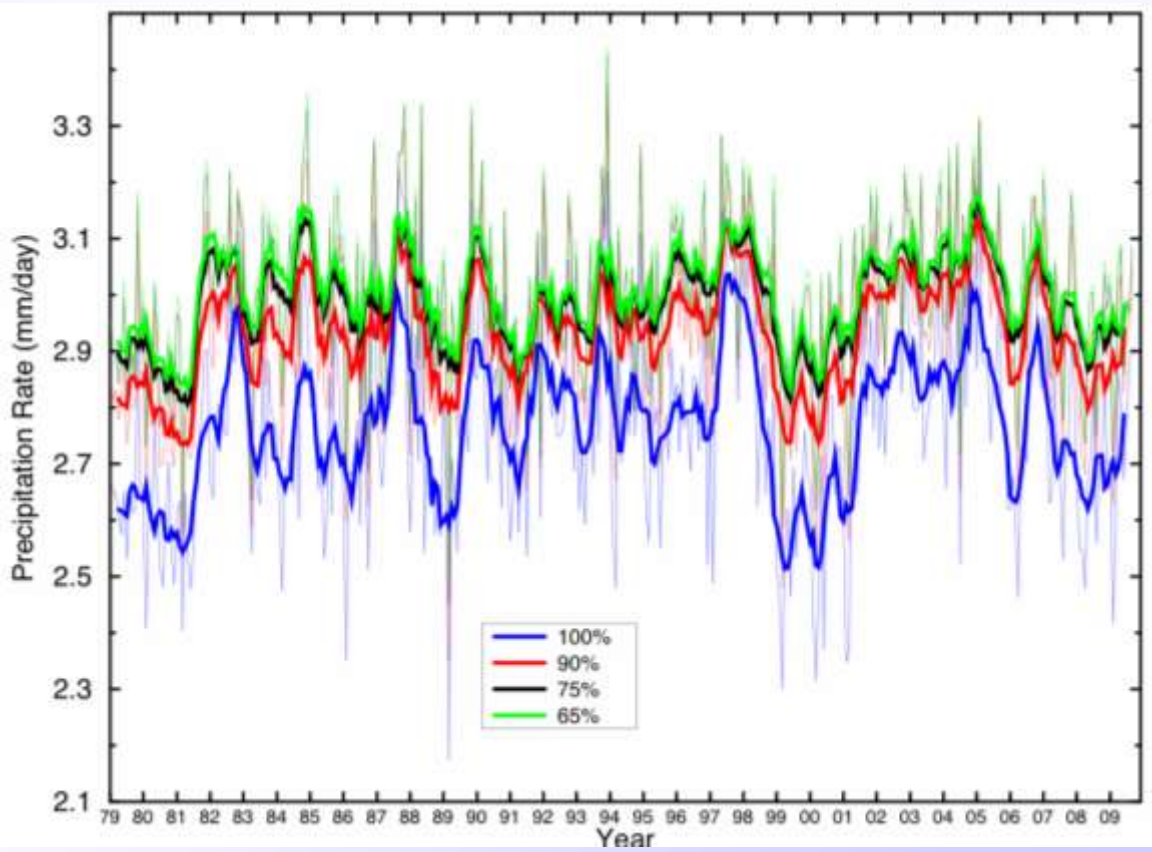
- the parameter of interest
- error estimates
  - bulk vs. gridded
  - RMS, bias, …
- ancillary data
  - dataset-specific (hot and cold loads, number of samples, …)
  - environmental (temp., humidity, wind, surface type)
  - a standard surface mask for land/coast/ocean can be important …

Other Considerations for CEWIS (cont.)

Which data? (cont.)

- example of “tropical ocean” time series for GPCP with different definitions of “ocean”
  - 2.5°grid
  - “fraction with sfc. water” from 100 to 65%
  - precip clusters at coast!
  - analyses with different definitions will differ





Summary

The devil is in the details

There are a lot of details

We've tended to do things ourselves

- control over choices
- more difficulty in making comparisons to other analyses
- more difficulty is maintaining our research momentum

Can CEWIS do some things well that make the rest of our work easier?